Deep Knowledge Group

Data Sources, Quality Assurance and Accuracy Overview
Longevity Investment Dashboard

May 2024

1. General Overview

1.1. **Deep Knowledge Group's** primary competitive edge lies in its analytical approach to sourcing data. In pursuit of this objective, we employ a diverse range of algorithms for collecting and processing data. Our team of experts is continuously innovating, creating new algorithms, and enhancing existing ones to further bolster our capabilities.

1.2. Our data collection process employs a range of sophisticated algorithms designed to efficiently gather information from diverse sources. These algorithms are tailored to specific data types and sources, enabling comprehensive data retrieval across various domains.

1.3. By utilising advanced parsing techniques and leveraging **APIs, NLP** and **web scraping** technologies, we ensure the accuracy and reliability of the collected data. These algorithms play a crucial role in enabling us to access, extract, and organise vast amounts of data, empowering us to derive valuable insights and make informed decisions.

2. Data Sources, Filling, Quality Assurance and Accuracy

Companies			
Datapoint	**Source**	**Data Quality**	**Data Filling**
name	Bing Answer + NER, biopharmguy.com	100%	100%
Super_Industrie	longevity company classificator (GTP promt respons)	85%	100%
Unique Industry	longevity company classificator (GTP promt respons)	85%	100%
Unique Sector	longevity company classificator (GTP promt respons)	85%	100%
headquarters	Bing Answer + NER, biopharmguy.com, Yahoo Finance	85%	83%
country	Bing Answer + NER, biopharmguy.com, Yahoo Finance	85%	83%
email	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	72%
website	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	97%
logo	Logo Parser, Logo Validator	90%	92%
description	Bing Answer + NER, biopharmguy.com, Yahoo Finance, Description parser	90%	95%

founders	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	25%
ceo	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	15%
employees_number	Bing Answer + NER, biopharmguy.com, Yahoo Finance	85%	37%
revenue	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	20%
investors	Bing Answer + NER, biopharmguy.com, Yahoo Finance	85%	12%
linkedin	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	43%
instagram	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	18%
facebook	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	43%
x	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	36%
IPO status	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	86%
Last Funding Date	Bing Answer + NER, biopharmguy.com, Yahoo Finance	80%	34%
Funding Status	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	31%
Company Type	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	78%
ticker	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	6%
founded_date	Bing Answer + NER, biopharmguy.com, Yahoo Finance	90%	72%

Investors

Datapoint	Source	Data Quality	Data Filling
name	Bing Answer + NER	100%	100%
logo	logo parser, logo validator, www.linkedin.com	90%	98%
Investor Type	investor classificator (GTP promt respons)	90%	81%
country	Bing Answer + NER (LOC entity)	87%	87%
company invested in	Bing Answer + NER (ORG entity)	85%	11%
description	Bing Answer + NER, website text extractor, description parser	90%	86%
headquarters	Bing Answer + NER (LOC entity)	85%	34%

Unique Industry	longevity classificator (GTP promt respons)	90%	86%
Unique Sector	longevity classificator (GTP promt respons)	90%	86%
website	Bing Answer, yahoo finance, website text extractor, www.linkedin.com	90%	63%
email	Bing Answer ,yahoo finance, website text extractor, www.linkedin.com	90%	67%
founders	Bing Answer + NER (PER), yahoo finance	87%	11%
employees_number	Bing Answer + NER (CAR), yahoo finance	85%	24%
revenue	Bing Answer + NER (CAR), yahoo finance	87%	8%
founded_date	Bing Answer + NER (DATE), yahoo finance	90%	74%
founded_year	Bing Answer + NER (DATE), yahoo finance	90%	73%
market_cap	Bing Answer + NER (CAR), yahoo finance	90%	74%
linkedin	Bing Answer + NER	90%	32%
facebook	Bing Answer, www.linkedin.com	90%	19%
x	Bing Answer, www.linkedin.com	90%	17%
Headquarters Location	Bing Answer + NER (LOC entity)	85%	85%

Hubs & Longevity Organizations

Datapoint	Source	Data Quality	Data Filling
name	Bing Answer + NER, biopharmguy.com	100%	100%
Type	HUBS classificator (GTP promt respons)	90%	100%
logo	logo parser, logo validator, www.linkedin.com	90%	95%
website	Bing Answer, yahoo finance, website text extractor, www.linkedin.com	90%	85%
email	Bing Answer ,yahoo finance, website text extractor, www.linkedin.com	90%	59%
headquarters	Bing Answer + NER (LOC entity)	85%	50%
country	Bing Answer + NER (LOC entity)	87%	50%
ceo	Bing Answer + NER (PER), yahoo finance	85%	13%
employees_number	Bing Answer + NER (CAR), yahoo finance	85%	45%
revenue	Bing Answer + NER (CAR), yahoo finance	87%	16%

founded_date	Bing Answer + NER (DATE), yahoo finance	90%	13%
founded_year	Bing Answer + NER (DATE), yahoo finance	90%	10%
description	Bing Answer + NER, website text extractor, description parser	90%	77%
founders	Bing Answer + NER (PER), yahoo finance	87%	17%
linkedin	Bing Answer + NER (CAR entity)	90%	48%
instagram	Bing Answer, www.linkedin.com	90%	23%
facebook	Bing Answer, www.linkedin.com	90%	39%
x	Bing Answer, www.linkedin.com	90%	32%

Leaders			
Datapoint	**Source**	**Data Quality**	**Data Filling**
full_name	Bing Answer + NER, Apollo Parser, Linkedin Parser	100%	100%
photo	Photo Parser, Photo Validator	90%	80%
email	Bing Answer + NER, Apollo Parser, Linkedin Parser	90%	65%
linkedin	Bing Answer + NER, Apollo Parser	90%	77%
country	Bing Answer + NER, Apollo Parser, Linkedin Parser	90%	83%
headline	Bing Answer + NER, Apollo Parser, Linkedin Parser	90%	98%
company	Bing Answer + NER, Apollo Parser, Linkedin Parser	90%	77%

3. Overall Data Filling, Quality Assurance and Accuracy

3.1. **Companies**. Average Data Quality Assurance and Accuracy is 89%. Average Data Filling Assurance and Accuracy is 66%.

3.2. **Investors**. Average Data Quality Assurance and Accuracy is 89%. Average Data Filling Assurance and Accuracy is 68%.

3.3. **Hubs & Longevity Organizations**. Average Data Quality Assurance and Accuracy is 89%. Average Data Filling Assurance and Accuracy is 60%.

3.4. **Leaders**. Average Data Quality Assurance and Accuracy is 91%. Average Data Filling Assurance and Accuracy is 85%.

3.5. **Acquisitions**. Average Data Quality Assurance and Accuracy is 85%. Average Data Filling Assurance and Accuracy is 75%.

3.6. **Funding Rounds**. Average Data Quality Assurance and Accuracy is 85%. Average Data Filling Assurance and Accuracy is 80%.

4. Data Sources

4.1. Bing Answer + NER (Named Entity Recognition) is a structured process for gathering specific data entities from various online sources mainly from google and bing. This methodology combines search engine capabilities and advanced entity recognition techniques to extract and verify relevant information. Bing Answer + NER provides a robust framework for collecting and verifying a wide range of data entities. By integrating search engine capabilities with advanced NER methods, this approach ensures comprehensive and reliable data extraction across various domains.

4.2. The process of collecting logos takes place by making a request to google, bing search engines and obtaining the URL of the image according to the entered information, downloading the image of the logo from the appropriate source or resource using the URL of the image, and checking whether the image corresponds to the required formats (eg JPEG, PNG, etc.). Validation of process images using NNtool validation classification.

5. Data Classification

5.1. At DKG, we specialise in advanced text preprocessing solutions that optimise data analysis processes for businesses across industries. Our expertise lies in refining raw text data, extracting key objects, and ensuring data integrity through thorough cleaning and structuring. Text pre-processing is an important step in the data analysis pipeline, as it lays the foundation for accurate and insightful conclusions. By cleansing the data and extracting relevant objects, we reduce noise and improve the quality of information, thereby contributing to more effective decision-making and strategic planning for our clients.

5.2. Our text pre-processing services offer several advantages:

5.2.1. Improved data quality: We eliminate inconsistencies, errors, and redundancies in text, resulting in cleaner, more reliable datasets.

5.2.2. Improved object recognition: With advanced algorithms and natural language processing techniques, we excel at identifying and extracting key objects such as dates, names, locations, organisations, and more, enriching the dataset with valuable contextual information.

5.3. **Optimized analysis**: By organising and structuring text according to predefined criteria, we simplify the data analysis process, saving time and resources while maximising the depth of insights gained from the data.

5.4. **Increased accuracy**: Our rigorous pre-processing methods ensure that data is properly formatted and standardised, minimising the risk of misinterpretation and errors during analysis.

5.5. **Entity recognition and extraction**: To extract named entities such as dates, names, locations, organisations, etc., we use models trained specifically for named entity recognition (NER). These models are typically based on deep learning architectures such as bidirectional LSTMs (long short-term memory networks) or transformer-based models such as BERTs (bidirectional encoder representations of transformers).

5.6. **GTP prompt classification**: Regarding the conditions of classification, it is assumed that the framework of the relation of classes, sub-classes, etc. is formed, the classification is one-way or multi-way. For each specific task of classification, the optimal GTP prompt is formed, and the corresponding filters are formed to form a structured answer.

6. Entity Recognition

6.1. Deep Knowledge Group (DKG) has spearheaded the development of advanced Entity Recognition algorithms, revolutionising our ability to extract validated data from vast quantities of unstructured sources. These algorithms represent a breakthrough in data processing, enabling us to efficiently identify and validate entities within text and images, thereby enhancing the quality and reliability of our dataset.

6.2. Our Entity Recognition algorithms are designed to tackle the inherent challenges posed by unstructured data, such as news articles, reports, social media posts, and other textual sources. Leveraging state-of-the-art natural language processing (NLP) techniques, machine learning models, and semantic analysis, these algorithms can accurately identify and extract entities, including entities such as organisations, people, locations, events, and more. By discerning context, relationships, and relevance, our algorithms ensure that extracted entities are not only identified but also validated for accuracy and reliability.

6.3. In addition to text, our Entity Recognition algorithms extend to image data, enabling us to validate and extract entities from visual content with precision and efficiency. Through advanced image recognition and pattern recognition techniques, we can identify objects, logos, landmarks, and other visual elements, enriching our dataset with valuable insights derived from multimedia sources.

6.4. Furthermore, our data enrichment and classification validation processes are augmented by querying search engines, leveraging their vast repositories of indexed information to corroborate and validate extracted entities. By cross-referencing data obtained from multiple sources, we enhance the accuracy and completeness of our dataset, ensuring that only validated information is incorporated into our database.

6.5. A key aspect of our Entity Recognition framework is the integration of a data validator, which provides continuous evaluation and feedback to the system. This feedback loop enables our algorithms to learn and adapt over time, incorporating new patterns, insights, and corrections to improve their accuracy and performance. By leveraging machine learning techniques, the system can dynamically adjust its algorithms based on real-world feedback, ensuring that the database remains clean, updated, and reflective of the most current information available.

7. Advantages of Utilising APIs, NLP and Web Scraping Technologies

7.1. **Efficiency**: By employing APIs, NLP, and web scraping, we optimise the data collection process, reducing manual efforts and enhancing efficiency. These technologies automate data retrieval tasks, allowing our team to focus on analysis and decision-making rather than data acquisition.

7.2. **Accuracy**: The use of advanced parsing techniques and structured data access through APIs ensures the accuracy and reliability of the collected data. NLP algorithms enhance data understanding and interpretation, enabling us to extract meaningful insights with precision.

7.3. **Comprehensive Data Coverage**: By combining APIs, NLP, and web scraping technologies, we achieve comprehensive data coverage across various sources and domains. This approach enables us to gather data from diverse sources, including structured databases and unstructured web content, facilitating holistic analysis and informed decision-making.

7.4. **Real-Time Access**: APIs provide real-time access to data, ensuring that our analyses are based on the latest information available. This real-time data access enhances the timeliness and relevance of our insights, enabling agile responses to changing market conditions and emerging trends.

7.5. **Flexibility**: The flexibility of web scraping allows us to extract data from any website, regardless of whether an API is available. This flexibility enables us to adapt to evolving data sources and requirements, ensuring continuity and adaptability in our data collection efforts.

7.6. **Innovation**: Deep Knowledge Group's commitment to innovation drives continuous improvement in our algorithms and technologies. By staying at the forefront of advancements in APIs, NLP, and web scraping, we continually enhance our capabilities and maintain a competitive edge in data analytics and decision support.

8. Data Quality Control, Accuracy and Lawfulness

8.1. Data quality control and accuracy are paramount in Deep Knowledge Group's (DKG) analytical endeavours, particularly given the diverse array of data sources we rely on. Our commitment to excellence begins with the careful selection of source material, drawn from a multitude of reputable outlets, including news articles, press releases, organisation websites, scientific publications, patent databases, government websites, and various open sources encompassing a wide spectrum of information.

8.2. Recognising the inherent variability and potential biases across different data sources, we prioritise independent data from external sources outside our organisation. This approach not only enriches our dataset with diverse perspectives but also serves as a foundational principle for ensuring the reliability and credibility of our analyses. By cross-validating each data point across multiple sources whenever possible, we mitigate the risk of inaccuracies and inconsistencies, thereby bolstering the overall quality and trustworthiness of our database.

8.3. The process of data validation and quality control at DKG is meticulous and thorough. We employ a multifaceted approach that involves rigorous scrutiny of each data point, leveraging advanced algorithms, machine learning techniques, and human expertise to assess its validity, relevance, and reliability. Through automated checks and manual verification procedures, we identify and weed out low-quality, outdated, or questionable data, ensuring that only the most accurate and up-to-date information is retained in our database.

8.4. Furthermore, we recognise that maintaining data accuracy is an ongoing endeavour that requires constant vigilance and proactive measures. As such, we have implemented robust protocols for continuous monitoring, validation, and updates, enabling us to promptly address any discrepancies, errors, or changes in the underlying data sources. Whether it's through regular audits, real-time alerts, or feedback mechanisms from users and stakeholders, we remain vigilant in our efforts to uphold the integrity and accuracy of our data.

8.5. In addition to stringent quality control measures, we also prioritise transparency and accountability in our data management practices. We provide clear documentation regarding the sources, methodologies, and assumptions underlying our analyses, allowing users to assess the reliability and credibility of the insights derived from our data. Moreover, we actively engage with stakeholders to solicit feedback, address concerns, and refine our data collection and validation processes, thereby fostering a culture of continuous improvement and trust.

8.6. From a **legal** standpoint, we rely on certain legal exclusions provided by various legislation in the UK, EU and US concerning data collection and mining. In particular, based on the jurisdiction and aims of data collection, its further use, we rely on temporary reproduction in the UK and EU, the text and data mining exception in the EU, and fair use in the US.

8.7. At each stage, specific legal criteria receive and filter each individual data source:

8.7.1. Access to the content: Initial access to the content is subject to stringent legal scrutiny to ensure compliance with relevant regulations and standards;

8.7.2. Extraction and copying of content: The process of extracting and copying content is conducted within the bounds of applicable laws, with careful attention paid to intellectual property rights and data protection considerations;

8.7.3. Text and data mining: Text and data mining activities are conducted in accordance with the provisions and exceptions afforded by relevant legislation, balancing the imperative of innovation with the imperative of safeguarding privacy and intellectual property rights;

8.7.4. Data storage: The storage of mined data is executed in compliance with established data protection regulations, encompassing measures aimed at safeguarding the confidentiality, integrity, and availability of the data.

8.8. Furthermore, our adherence to privacy regulations is unwavering, with a particular focus on the EU General Data Protection Regulation, the UK General Data Protection Regulation, and the Data Protection Act 2018. With each occurrence of personal data collection, we establish legal grounds and implement a comprehensive array of supplementary measures. For instance, we provide users with expedited mechanisms such as quick contact forms for the deletion or rectification of personal data, ensuring that individuals retain agency over their personal information within the parameters of the law.

8.9. We diligently observe and abide by AI regulations promulgated by regulatory bodies such as the European Commission and the UK government. Specifically, we adhere to the EU Regulation on Artificial Intelligence (AI Act) and the UK's AI Regulatory Framework, which provide comprehensive frameworks for the ethical and lawful deployment of AI technologies, including those involved in data collection and mining endeavours.

8.10. Our AI regulatory compliance efforts encompass several key facets:

8.11. **Transparency and Accountability**: We prioritise transparency in AI-driven data collection processes, ensuring that individuals are informed about the nature, scope, and implications of data collection activities. Moreover, we uphold principles of accountability, assuming responsibility for the ethical and lawful utilisation of AI technologies in data mining endeavours.

8.12. **Fairness and Non-discrimination**: In accordance with regulatory mandates, we uphold principles of fairness and non-discrimination in AI-driven data collection practices. We employ measures to mitigate biases and ensure equitable treatment across diverse demographic groups, thereby fostering trust and inclusivity in our data collection initiatives.

8.13. **Data Protection and Privacy**: As indicated above we prioritise the safeguarding of individuals' privacy rights and personal data integrity throughout the data collection lifecycle.

8.14. **Ethical Considerations**: We remain attuned to ethical considerations inherent in AI-driven data collection and mining activities, striving to uphold principles of beneficence, autonomy, and justice in our practices. Our ethical framework guides decision-making processes, ensuring alignment with societal values and ethical norms.